ANALYSIS

45(209), October 1, 2015

ISSN 2278–5469
EISSN 2278–5450

Discovery

Pattern based approach for mining users opinion from Kannada web documents

Anil Kumar KM[1], Asmita Poojari[2], Mohana kumari M[3]

1.Associate Professor, Dept.of CS & E, Sri Jayachamarajendra College of Engineering, Mysore, Karnataka, India; Email: anilkmsjce@yahoo.co.in
2.Assistant Professor, Dept. of CS & E, NMAMIT, Nitte, Karnataka, India
3.Dept. of CS & E, NMAMIT, Nitte, Karnataka, India; Email: mohana.razz@gmail.com

Article History
Received: 01 August 2015
Accepted: 18 September 2015
Published: 1 October 2015

Citation
Anil Kumar KM, Asmita Poojari, Mohana kumari M. Pattern based Approach for Mining Users Opinion from Kannada Web Documents. *Discovery*, 2015, 45(209), 138-143

ABSTRACT

Opinions are very important for any organization's development and success as it affects the sale of its products. These opinions have impact on the customers buying decisions related to various products. With popularity of cheaper smart phones, access to internet and web pages has become easier. Nowadays Kannada websites are very popular for Kannada web users and they rely on these websites for accessing their day today information. One such information they access is about opinion of others related to variety of topics. Hence opinions are very useful for many web users, such as business analysts, directors and producers of movies etc. In this paper we present a pattern based approach for effective mining user’s opinion from Kannada web documents. We find that our pattern based approach provides better results than the other pattern discussed in literature.

1. INTRODUCTION

Opinion mining is a process for tracking the mood of the public about a particular product or topic. Opinion mining is also called as Sentiment analysis [1] and it refers to the use of natural language processing, text analysis and computation linguistics. It is process of extraction of web user's feelings expressed in reviews i.e. positive or negative, through the analysis of a large number of documents. In general sentiment analysis aims to determine the opinion of a web user with respect to the topic. In recent years, many people use the internet to share their opinions on products, services, movies, restaurants, vacation places etc. Social media is a good platform for sharing their opinions about anything and it can be accessible by anybody on the web. These opinions are very useful for many web users. Today a large amount of information is available on the web about products or services. Some of these reviews are very long and making it impossible for user to be patient enough to read the complete review. In these cases, web users resort to reading short reviews or part of these reviews in order to take decision regarding the products or services. There are many approaches that are available for the classification of sentiment (positive sentiment or negative sentiment) in English language. Also there are no studies undertaken to find feasibility of existing approaches related to English language on Indian languages. Many people's share their opinions in native languages across the websites, for this reason there is a need for sentiment classification in native languages. In this paper we aim to detect opinions from Kannada text. For example, "ನೊಕಿಯಾ ಮೊಬೈಲ್ ತುಂಬಾ ಚೆನ್ನಾಗಿದೆ." Instead of "Nokia is a very good phone", these type of reviews are very useful for native language speaker such as Kannada. In this paper, we check the feasibility of well known extraction pattern of English language to Kannada language and propose a set of patterns that can effectively detect user's sentiment from Kannada web documents.

2. RELATED WORKS

2.1. Sentiment Analysis

The analysis of data to extract latent public opinion and sentiment is a challenging task. Liu et al. [2] defines a sentiment or opinion as a quintuple-

"<oj, fjk, soijkl, hi, tl >, where oj is a target object, fjk is a feature of the object oj, soijkl is the sentiment value of the opinion of the opinion holder hi on feature fjk of object oj at time tl, soijkl is +ve,-ve, or neutral, or a more granular rating, hi is an opinion holder, tl is the time when the opinion is expressed."

Pang-Lee et al [3] broadly classifies the applications into the following categories.

- Applications to Review-Related Websites, Movie Reviews, Product Reviews etc.
- Applications as a Sub-Component Technology Detecting antagonistic, spam detection, context sensitive information detection etc.
- Applications in Business and Government Intelligence, Knowing Consumer attitudes and trends
- Applications across Different Domains Knowing public opinions.

According to Collomb et al. [4], the existing work on sentiment analysis can be classified from different points of views: technique used, view of the text, level of detail of text analysis, rating level, etc.

2.2. Part-of-speech Tagging in Kannada

According to Vijayalaxmi F Patil [5], in most of the Dravidian languages, particularly for Kannada language, nouns and verbs get inflected. Also verbs and adjectives are nominalized by means of certain nominalizers. Siva Reddy and Serge Sharoff [6] have built a Hidden-Markov model (HMM) based Kannada POS tagger. The tagset has both POS and morphological information encoded in it, the HMM model has an advantage of using morphological information to predict the main POS tag, and the inverse, where main POS tag helps to predict the morphological information. Akshar Bharati et al. [7] arrive at standard tagging scheme for POS tagging and chunking for annotating Indian languages (AnnCorra) and come up with the tags which are exhaustive for the task of annotation for Indian languages.

2.3. Turney's method

Turney et al. [8] present a simple unsupervised learning algorithm for classifying a review. The algorithm takes a written review as input and produces a speech classification as output. The first step is to use a part-of-tagger to identify phrases in the input text that contain adjectives or adverbs. The second step is to estimate the semantic orientation of each extracted phrase. The third step is to assign the given review to a class, recommended or not recommended, based on the average semantic orientation of the phrases extracted from the review. If the average is positive, the prediction is that the review recommends the item it discusses. Two consecutive words are extracted from the review if their tags conform to any of the patterns in Table 1. The JJ tags indicate adjectives, the NN tags are nouns, the RB tags are adverbs, and the VB tags are verbs. The second pattern, for example, means that two consecutive words are extracted if the first word is an adverb and the second word is an adjective, but the third word (which is not extracted) cannot be a noun. NNP and NNPS (singular and plural proper nouns) are avoided, so that the names of the items in the review cannot influence the classification. The second step is to estimate the semantic orientation of the extracted phrases, using the PMI-IR algorithm. This algorithm uses mutual information as a measure of the strength of semantic association between two words.

2.4. Sentence based approach

Khan and Baharudin [9] discuss about sentiment analysis at individual sentence level in which from subjective sentences, the opinion expressions are extracted and their semantic scores are checked using the SentiWordNet directory. The final weight of each individual sentence is calculated after considering the whole sentence structure.

3. METHODOLOGY

In the present scenario there is no published work for classifying sentiments expressed in Kannada language. For this purpose, we develop certain algorithms, applicable to document and sentences based approaches to analyze the sentiment expressed in Kannada. First, we start by creating Kannada dictionary using sentiments from the English dictionary [10] [11]. To develop the dictionary, we follow a manual approach to identify sentiment words from the reviews and also use translation of English keywords into Kannada using Google translator. We also build a list of negators to capture words that reverses the polarity of a sentiment word. The English dictionary has 2006 positive words and 4784 negative words. Similarly, Kannada dictionary has 1302 positive words and 1789 negative words. The negators used in English language are no, not and don't. Similarly, negators used in Kannada language are ಇಲ್ಲ , ಇಲ್ಲಾ , ಪಡುವುದಿಲ್ಲ , ಸುಲಭವಲ್ಲ , ಅಲ್ಲ , ಆಗಿಲ್ಲ , ಆಗಲ್ಲಾ , ಆಗಲಿಲ್ಲ , ಬರುತ್ತಿಲ್ಲ , ಹೇಳಕ್ಕಾಗಲ್ಲ , ಖರೀದಿಯಲ್ಲ , ಹೊಂದಿರುವುದಿಲ್ಲ , ಆಗುವುದಿಲ್ಲ , ಮಾಡುವುದಿಲ್ಲ , ಕೊಡುವುದಿಲ್ಲ , ಸಿಗುತ್ತಿಲ್ಲ and ನೀಡುತ್ತಿಲ್ಲ. We make use of Monty Tagger [12] and Kannada POS Tagger [13] software for extracting patterns useful for detecting sentiments. Monty Tagger is a popular natural language processing toolkit. For English sentences, it extracts subject/verb/object tuples, adjectives phrases, noun phrases, verb phrases, and other semantic information. For example, consider an opinionated text "great camera canon I bought my canon g3 about a month ago and i have to say i am very satisfied." After passing it through a parts of speech program, we obtain the tagged opinionated sentences such as great/JJ camera/NN canon/NN i/NN bought/VBD my/PRP$ canon/NN g3/CD about/IN a/DT month/NN ago/RB and/CC i/NN have/VBP to/TO say/VB i/NN am/VBP very/RB satisfied/VBN ./. Where JJ represent adjective, NN represents noun, CD represent cardinal, RB represents adverb and VBP/VBN/VBD represent different forms of verb. Similarly, Kannada POS Tagger developed by Siva Reddy and Serge Sharoff can be used for analysing and annotating Kannada texts. For example, ನನಗೆ ಜಾನ್ಸನ್ ಬೇಬಿ ಕ್ರೀಮ್ ತುಂಬಾ ಇಷ್ಟ ಯಾಕೆ ಅಂದ್ರೆ ಅದು ಎಣ್ಣೆ ಎಣ್ಣೆಯಾಗಿ ಮುಖಾನ ಮಾಯಿಶ್ಚರೈಸರಿಂದ ಕಾಪಾಡುತ್ತೆ. After passing it through a splitter program and POS tagger program, we obtain the tagged sentences such as

ನನಗೆ	PRP.pn..sg.1.o	ನಾನು.ಇಗೆ
ಜಾನ್ಸನ್	NNP.unk....	ಜಾನ್ಸನ್
ಬೇಬಿ	NN.unk....	ಬೇಬಿ.
ಕ್ರೀಮ್	NN.unk....	ಕ್ರೀಮ್.
ತುಂಬಾ	INTF.avy....	ತುಂಬಾ.
ಇಷ್ಟ	JJ.n.n.sg..d	ಇಷ್ಟ.0
ಯಾಕೆ	WQ.avy....	ಯಾಕೆ.
ಅಂದ್ರೆ	VM.unk....	ಅಂದ್ರೆ.
ಅದು	PRP.n.n.sg..d	ಅದು.0
ಎಣ್ಣೆ	NN.n.n.sg..d	ಎಣ್ಣೆ.0
ಎಣ್ಣೆಯಾಗಿ	RB.unk....	ಎಣ್ಣೆಯಾಗಿ.
ಮುಖಾನ	NN.UNK....	ಮುಖಾನ.
ಮಾಯಿಶ್ಚರೈಸರಿಂದ	NN.UNK....	ಮಾಯಿಶ್ಚರೈಸರಿಂದ.
ಕಾಪಾಡುತ್ತೆ	VM.unk....	ಕಾಪಾಡುತ್ತೆ.

Table 1 and Table 2 shows extraction patterns used to find opinionated words in English and Kannada language. We use the pattern as discussed in Table 1 to detect sentiments at document level and sentence level. For example, "This is an excellent phone! Lives up to the expectation." After applying pos tagging, we obtain This/DT is/VBZ an/DT excellent/JJ phone/NN !/. Lives/NNS up/IN to/TO the/DT expectation/NN. /. When the extraction patterns are applied, we obtain excellent/JJ phone/NN as opinionated words. If the word is found in positive dictionary, then the word is considered as positive and assigned score +1. If the word is found in negative dictionary, then the word is considered as negative and assigned score -1

Table 1 Turney's patterns [8]

Slno	First word	Second word	Third word
1	JJ	NN or NNS	Anything
2	RB, RBR, or RBS	JJ	Not NN nor NNS
3	JJ	JJ	Not NN nor NNS
4	NN or NNS	JJ	Not NN nor NNS
2	RB, RBR, or RBS	VB,VBD,VBN or VBG	anything

Table 2 Our Extraction patterns

Slno	First word	Second word
1	JJ	NN
2	INTF	VM,JJ,RB or NN
3	NN	VM
4	RB	VM
5	PRP	JJ or VM

After finding the polarity of the words, the opinionated sentence is parsed to find negators at window size of 3 from the occurrence of the opinionated words, since we believe negators appear in closer proximity to opinionated words. If the negators are present with positive word, positive score is multiplied by -1 and score become -1, then opinionated word is considered as negative. Similarly if negators are present with negative word, negative score is multiplied by -1 and score become +1, implying opinionated word to be positive. If there is no negator in an opinionated text, the score of the opinionated text will be similar to the evaluated score of opinionated words. We compute the average semantic orientation of the opinionated text by considering all scores of opinionated words shown in Equation 1. Here, we have used threshold value as 0. If the average semantic orientation of opinionated text is greater than threshold value, then opinionated text is positive. If it is less than threshold value, then it is negative and if it is equal to zero, then it is neutral.

$$\text{SO (Opinionated Text)} = \sum_{i=1}^{n} \text{(Opinionated words i)} \qquad (1)$$

We employ the following algorithms for detecting users opinion from texts.

Algorithm 1: Turney's algorithm

In this algorithm, we try to recognise various patterns formed by the words as developed by Turney [8] and when a pattern is found, we compute the polarity of the words.

```
Tag the text file using POS tagger
   For each words in the file
-If the word matches any of the Turney's   patterns     compute the polarity by analyzing the words and SO(text)
      End if
End for
```

Algorithm 2: Negator algorithm

This algorithm is an enhancement to the above algorithm negator algorithm uses the list of negators which reverse the polarity of the keyword. The algorithm searches a negator for each encountered opinionated word within the window size 3 i.e. three words before the opinionated word in case of English language and three words after the opinionated word in case of Kannada language. The negators are stored in a separated file.

```
If the words matches any of the patterns
  If the word is found in positive/negative list
   For each word within the window
        If negator is found
          Multiply positive/negative score by -1
 Compute SO (text)
        End if
    End For
  End if
End if
```

We have considered the sentence based approach. Sometimes people express their opinion in first line or last line of the document. We will apply the above algorithm to the sentence level and analyze the results.

Algorithm 3: Significant Sentence algorithm

```
Tag the file using POS tagger
Split the document into separated lines
For each sentence in the file
      Select the sentence
     Compute SO (text)
End for
Compute the polarity of the sentence
```

4. EXPERIMENTS & RESULTS

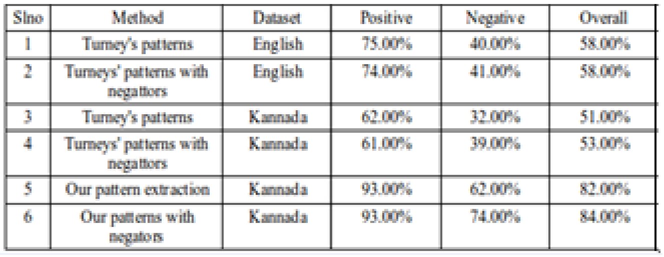
In this paper, as a part of the experiment, we have collected 183 positive Kannada reviews, 105 negative Kannada reviews, 200 positive English reviews and 180 negative English reviews as our data set. We consider 3 special cases in sentence approach i.e. first sentence, last sentence and first and last sentence as significant sentences to find the opinion. The reviews were mainly collected from broad domains consisting of commercial products like automobiles, health and body care products like soaps, shampoo, electronic items like TV, mobiles, movies, songs, websites, TV programs, etc. To validate the words in the Kannada dictionary, we used 300 evaluators to find out the polarity of words with 60 % agreement. i.e. word is treated as positive word or negative or neutral based on agreement. Finally, it consists of 1302 positive words and 1789 negative words for evaluating opinionated words. All the Kannada/English reviews were stored as separate files in two separate folders one each for positive set and negative set. The programs were designed to take review input as file from the folders. All the Kannada reviews were entered in UTF-8 format. We used Standard Kannada POS tagger software for Kannada language and Monty tagger software for English language.

The algorithms which we defined earlier were applied to the dataset and obtained the following results: For Document-level approaches, we applied patterns as prescribed by Turney's algorithm for English dataset and obtained an accuracy of 58%. Also we obtain an accuracy of 58% considering Turney's algorithm with Negator. We compute an accuracy using Equation 2. The detailed result of this algorithm is as shown in Table 3. Similarly, we applied patterns as prescribed by Turney's algorithm for Kannada dataset and obtained an accuracy of 51%. Also we obtain an accuracy of 53% considering Turney's algorithm with negator. We also applied our new patterns, as shown in Table 2, for Kannada dataset and obtained an accuracy of 82%. Also we obtain an accuracy of 84% considering Pattern extraction algorithm with negator. The detailed result of this algorithm is as shown in Table 3.

$$\text{Accuracy} = \frac{\text{number of correctly classified opinions in text document}}{\text{total number of opininated text files}} \quad (2)$$

For Sentence-level approaches, we considered three cases i.e. First sentence, Last-sentence and First and last-sentences. We applied patterns as described by Turney's algorithm and our patterns with different algorithm for English dataset and Kannada dataset. The detailed result of the entire algorithm is as shown in Table 4. For example, in English language and opinion is expressed as "it is good phone", with opinionated word appearing first followed by the subject. In Kannada, an opinion is expressed as "ನೊಕಿಯಾ ಮೊಬೈಲ್ ತುಂಬಾ ಚೆನ್ನಾಗಿದೆ" with subject appearing first followed by opinionated word. Hence considering these structure difference of language, new patterns for Kannada language were conceived to provide better opinion detection.

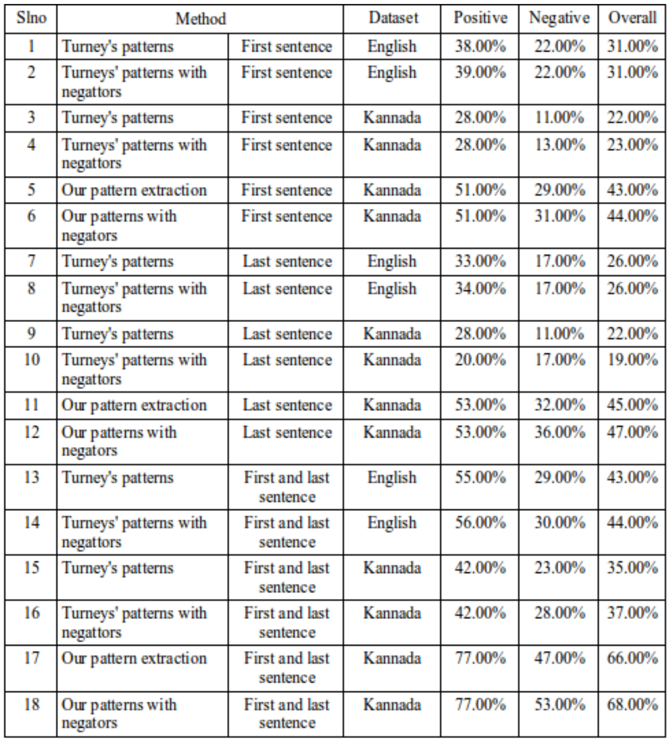
Table 3 Document approach



Slno	Method	Dataset	Positive	Negative	Overall
1	Turney's patterns	English	75.00%	40.00%	58.00%
2	Turneys' patterns with negattors	English	74.00%	41.00%	58.00%
3	Turney's patterns	Kannada	62.00%	32.00%	51.00%
4	Turneys' patterns with negattors	Kannada	61.00%	39.00%	53.00%
5	Our pattern extraction	Kannada	93.00%	62.00%	82.00%
6	Our patterns with negators	Kannada	93.00%	74.00%	84.00%

We have applied the algorithm for the review classification in both Kannada and English review sets and obtained the above results. We have found out that Pattern extraction algorithm with negator in Kannada with document based approach outperforms all other approaches and it provides good results than the extraction pattern of Turney's algorithm. In sentence based approach, first and last sentence were more accurate than first and last sentences and also we found that classification of positive reviews were more accurate than classification of negative reviews. Finally we observed, Document based approach is more accurate than the Sentence based approaches.

Table 4 Sentence approach



Slno	Method		Dataset	Positive	Negative	Overall
1	Turney's patterns	First sentence	English	38.00%	22.00%	31.00%
2	Turneys' patterns with negattors	First sentence	English	39.00%	22.00%	31.00%
3	Turney's patterns	First sentence	Kannada	28.00%	11.00%	22.00%
4	Turneys' patterns with negattors	First sentence	Kannada	28.00%	13.00%	23.00%
5	Our pattern extraction	First sentence	Kannada	51.00%	29.00%	43.00%
6	Our patterns with negators	First sentence	Kannada	51.00%	31.00%	44.00%
7	Turney's patterns	Last sentence	English	33.00%	17.00%	26.00%
8	Turneys' patterns with negattors	Last sentence	English	34.00%	17.00%	26.00%
9	Turney's patterns	Last sentence	Kannada	28.00%	11.00%	22.00%
10	Turneys' patterns with negattors	Last sentence	Kannada	20.00%	17.00%	19.00%
11	Our pattern extraction	Last sentence	Kannada	53.00%	32.00%	45.00%
12	Our patterns with negators	Last sentence	Kannada	53.00%	36.00%	47.00%
13	Turney's patterns	First and last sentence	English	55.00%	29.00%	43.00%
14	Turneys' patterns with negattors	First and last sentence	English	56.00%	30.00%	44.00%
15	Turney's patterns	First and last sentence	Kannada	42.00%	23.00%	35.00%
16	Turneys' patterns with negattors	First and last sentence	Kannada	42.00%	28.00%	37.00%
17	Our pattern extraction	First and last sentence	Kannada	77.00%	47.00%	66.00%
18	Our patterns with negators	First and last sentence	Kannada	77.00%	53.00%	68.00%

5. CONCLUSION

There are many approaches to classify the sentiments in English language. Also there are no studies undertaken to find feasibility of existing approaches related to English language to Indian languages. Here, we have discussed different extraction patterns applicable to detect opinion from Kannada opinionated text. We applied popular Turney's patterns for Kannada dataset and achieved accuracy of 51% and with negators 53%. We have developed new patterns for Kannada language and their application to Kannada text provides a better accuracy as against the original Turney's patterns. An increased accuracy of 33% and 31% is obtained with our new patterns as compared to Turney's patterns in the document based approach. Also, in the case of sentence based approach, we achieve an accuracy of 68% with first and last sentences. Our patterns have provided a better way to find opinions from collection of Kannada opinionated texts.

REFERENCE

1. http://en.wikipedia.org/wiki/Sentiment_analysis
2. Ramanathan Narayanan, Bing Liu and Alok Choudhary, Sentiment Analysis of Conditional Sentences, In Proceedings of Conference on Empirical Methods in Natural Language Processing, 2009
3. B. Pang and L. Lee, "Opinion mining and sentiment analysis," Foundations and trends in information retrieval, vol. 2, no. 1-2, pp. 1–135, 2008
4. A Study and Comparison of Sentiment Analysis Methods for Reputation Evaluation by Anaïs Collomb, Crina Costea, Damien Joyeux, Omar Hasan and Lionel Brunie
5. Designing POS Tagset for Kannada by Vijayalaxmi F. Patil, LDC-IL, CIIL Mysore
6. Cross Language POS Taggers (and other Tools) for Indian Languages: An Experiment with Kannada using Telugu Resources by Siva Reddy and Serge Sharoff

7. AnnCorra : Annotating Corpora Guidelines For POS And Chunk Annotation For Indian Languages by Akshar Bharati, Dipti Misra Sharma, Lakshmi Bai, Rajeev Sangal, Language Technologies Research Centre IIIT, Hyderabad
8. P. Turney, Thumbs up or thumbs down? Semantic Orientation Applied to Unsupervised Classification of Reviews, In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL'02), 2002
9. Sentiment Classification by Sentence Level Semantic Orientation using SentiWordNet from Online Reviews and Blogs by Aurangzeb khan and Baharum Baharudin, Universiti Teknologi PETRONAS Perak, Malaysia
10. https://github.com/jeffreybreen/twitter-sentiment-analysis-tutorial-201107/blob/master/data/opinion-lexicon-English/positive-words.txt
11. https://github.com/jeffreybreen/twitter-sentiment-analysis-tutorial- 201107/blob/master/data/opinion-lexicon-English/negative-words.txt
12. Hugo:MontyLingua: An end-to-end natural language processor with common sense. (2003)
13. http://sivareddy.in/downloads
14. SIMATIC NET PROFIBUS Networks manual, 6GK1970-5CA20-0AA1 Release 2 05/2000.
15. Dominique chabauty, "PROFIBUS Design and good practices".